

METHODOLOGY FOR ASSESSING SCALABILITY AND OPTIMIZING THE USAGE OF TOUGH2-MP ON A CLUSTER – APPLICATION CASE FOR A RADIOACTIVE WASTE REPOSITORY

Nicolas Hubschwerlen¹, Keni Zhang², Gerhard Mayer¹, Jean Roger³, Bernard Vialay³

¹AF-Consult Switzerland Ltd
Täferstrasse 26, 5405 Baden, Switzerland
e-mail: nicolas.hubschwerlen@afconsult.com

²College of Water Sciences, Beijing Normal University
19 Xijiekouwai St, Beijing, China

³Andra, Agence nationale pour la gestion des déchets radioactifs
Parc de la Croix Blanche, 92298 Châtenay-Malabry, France

ABSTRACT

Many numerical simulations are required to assess what impact heat and gas generated in the emplacement areas of a deep geological repository for radioactive waste may have over time on fluid pressure and saturation fields in the repository's drifts, shafts, and host rock. Ever-larger problems are being simulated, due to higher mesh resolutions and the consideration of larger scales (such as the full repository scale). To take advantage of cluster architecture, physical processes can now be simulated with the EOS5 module of the massively parallel multiphase flow and transport simulator TOUGH2-MP. However, the high demand in CPU time that such simulations require still makes optimal use of available computing resources a key issue.

Many TOUGH2-MP users may have little knowledge of, or experience in, how to efficiently set up, within their computational system, the parallel environment parameters of a TOUGH2-MP simulation. We have developed a new methodology whose purpose is to facilitate the efficient resource use of a cluster, by guiding the selection of the domain partitioning settings for the model and the distribution of the computation load among the nodes and cores.

This methodology involves a series of numerical test routines—such as parameter sensitivity analyses—which can be performed easily prior to the realization of production simulations.

The methodology is presented in this article using a simple case study based on the Couplex-Gaz 1b exercise, modeling the resaturation of a disposal cell for intermediate-level long-lived radioactive waste (ILW-LL), from its closure to the end of the gas production phase. The results demonstrate good scalability of TOUGH2-MP under certain conditions, as well as possible good practices for processor load distribution and the selection of the domain partitioning method. Parallel acceleration occurs with as little as (approximately) one thousand unknowns per partition, but optimal efficiency is achieved for larger partitions.

Finally, the methodology was validated successfully by applying it to a larger scale simulation case.

INTRODUCTION

The French National Agency for radioactive waste management (Andra) is currently investigating the feasibility of deep geological disposal of radioactive waste in an argillaceous formation (Andra, 2005). The long-term safety performance of this repository is strongly dependent on the impact that the heat and gas generated in the emplacement areas may have on the evolution of fluid pressure and saturation fields in repository drifts and shafts, as well as in the host rock itself.

In this context, many simulations of nonisothermal, multiphase fluid flow and multi-component transport in porous media are

performed using TOUGH2 and TOUGH2-MP (Pruess et al., 1999; Zhang et al., 2008). These simulations lead to very large computational demands that fully justify the use of the massively parallel TOUGH2-MP on a cluster. Trials on Andra's cluster Azurite showed that optimal use of TOUGH2-MP, which means achieving linear or even supralinear acceleration, is not that easy to predict using simple rules linking number of unknowns and number of parallel processes.

In this paper we present a methodology by which TOUGH2-MP on large clusters such as Azurite may be optimized. It was applied to two models of repository-relevant physical processes, material complexities and dimensions: first, the Couplex-Gaz 1b simulation model (Andra, 2006) and some derivate calculations; and second, a larger 3D model representing an entire ILW-LL emplacement cell.

METHODOLOGY

Performance objective and relevant parameters

In the current context of TOUGH2-MP usage, the objective is to perform many different calculations on a cluster where computational resources are shared and limited. Therefore, the chosen performance measure for optimal use is defined as efficiency in terms of the number of reference calculations performed per time and per utilized node¹. This performance measure allows some very easy comparisons between different computation strategies, in particular regarding node load and distribution of multiple computations on nodes. Moreover, this measure is very simple to obtain, since it relies exclusively on the measured CPU times provided by standard outputs of TOUGH2-MP and parallel computation parameters.

In the present work, performance dependency on hardware specifics was not considered, since all testing computations were performed on Andra's Azurite cluster. This equipment consists of 128

cores grouped in 16 computation nodes (2 quad-core Intel Xeon 5365 processors, 3GHz, 32GB RAM), for a peak accessible performance of about 1.5 Tflops. Dependency on the type of simulation problem was not studied either; the work was limited to one category of TOUGH2-MP EOS5 application: gas generation.

The study focused on the setup of parallel computations in terms of domain decomposition and repartition of the parallel processes on the cluster. Domain decomposition within TOUGH2-MP is performed using one of the METIS partitioner domain decomposition algorithms (Karypsis and Kummar, 1998, 1999). With respect to the repartition of parallel processes, we studied the influence of the cluster's nodes' loads.

Building a test case

The methodology requires a testing problem that must be as representative as possible of the real simulation cases performed with TOUGH2-MP EOS5 for the phenomenological analysis of a radioactive waste repository. This includes the modeled scenario, the discretization, and the choice of the numerical parameters. The number of unknowns typically ranges from a few thousand to several hundred thousand, and computation times may reach up to several days.

The representativeness of a test problem is ensured by respecting the following conditions:

- The modeled physical processes and parameters must be similar to those of the targeted real case. Some numerical parameters, such as solver parameters whose influence are not studied, are also considered as invariants and chosen as representative of the targeted case's settings.
- The computation must be able to perform several tests in a reasonable time. However, using the CPU time measurements from the TOUGH2-MP outputs requires long enough computation times to make measurement bias negligible.
- The test model must be large enough to profit from the parallel acceleration, i.e., its size must be representative with respect to the targeted problem as well.

¹ For the sake of legibility, and due to the average duration of the considered test runs, it is expressed in "runs performed per 1000 seconds per node".

- To study the scalability, i.e., estimate how well parallel acceleration results of TOUGH2-MP can be scaled to larger problems and numbers of partitions, the number of unknowns must be easy to vary, and this variation must be as neutral as possible on the convergence. A proposed solution is to use a pseudo-2D model and extend it by stacking several layers of it (Hubschwerlen et al., 2012).

One solution for limiting CPU time, while keeping a model large enough for parallel acceleration, is to limit the simulated time to a critical phase of a full simulation. One such phase occurs when saturation and pressure vary most, with very strong capillary pressure variations and variable production kinetics.

Partitioning, job repartition and scalability testing

On the basis of a test model built according to the rules set above, it is possible to test the computing performance behavior of TOUGH2-MP. Tests are performed in two steps. First, the impact of the different partitioning algorithms is investigated. Second, tests concerning the optimal number of parallel processes and node load repartition are conducted.

METIS provides three partitioning algorithms: *Recursive*, *Kway* and *VKway* (Karypis and Kumar, 1998). The TOUGH2-MP documentation suggests that the optimal METIS algorithm choice depends on the number of domain partitions (Zhang et al., 2008). Therefore, the performed test will consist of comparing, for the three available algorithms, the performance of a TOUGH2-MP calculation when varying the number of subdomains. Moreover, the model must be selected in order to have large-enough subdomains to benefit from parallel acceleration, even with the greatest number of nodes available. The (lower) threshold is estimated to be 1000 equations per parallel process (Hubschwerlen et al. 2012).

Once the optimal partitioning algorithm has been determined, some tests on the node load are performed, by varying the number of parallel processes being simultaneously run on one node

of the cluster, from one to the maximal load per node allowed by the cluster resources manager. In the case of Azurite, this maximum was set to 8 processes per nodes, which coincides with the number of cores. This setup is, however, not a general rule.

Varying the node load is achieved under invariant partitioning and invariant node repartition to ensure comparability of the different tests. This is achieved by running calculations with n parallel processes, each on a different node of the cluster, and augmenting the node load by raising the number of computations being run simultaneously (Figure 1).

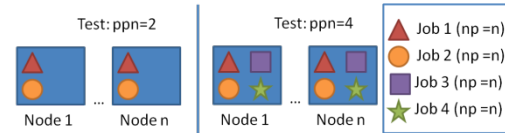


Figure 1. Scheme of node load variation test; np = nb. parallel processes per TOUGH2-MP run, ppn = nb. of processes per node.

A supplementary test to assess influence of node repartition is made by running simultaneously n times the same problem with the same decomposition (n subdomains and the same METIS algorithm), at constant node load, with all processes of one single run being dispatched alternatively on one to n nodes (Figure 2).

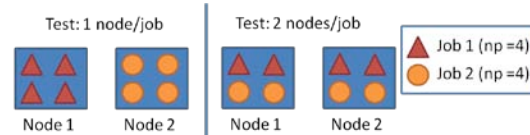


Figure 2. Scheme of node repartition test.

Finally, a scalability test is made by varying the number of unknowns of the test problem, to see if results obtained with a small test problem are representative for larger scale problems.

SCALABILITY TESTS BASED ON THE COUPLEX-GAZ 1B EXAMPLE

Preparation of the Couplex-Gaz 1b case for the tests

The methodology described above was applied on the Couplex-Gaz 1b case, which models the resaturation of a disposal cell for ILW-LL waste from its closure to the end of the gas production phase (Andra, 2006). The goal of the simulation

is to predict the position of the saturation front, the position of the dissolved-hydrogen front, the saturation variation in the different material zones of the repository, and the evolution of the gas and water pressure with time. TOUGH2-MP EOS5 is used with isothermal conditions (two equations per element).

The Couplex-Gaz 1b problem is an ideal candidate for applying the developed methodology, since highly heterogeneous materials and capillary properties make the problem very demanding in terms of computation needs. Moreover, the model layout has only 2480 elements on one layer, which makes it quite suitable as an elementary mesh for scalability testing (Figure 3).

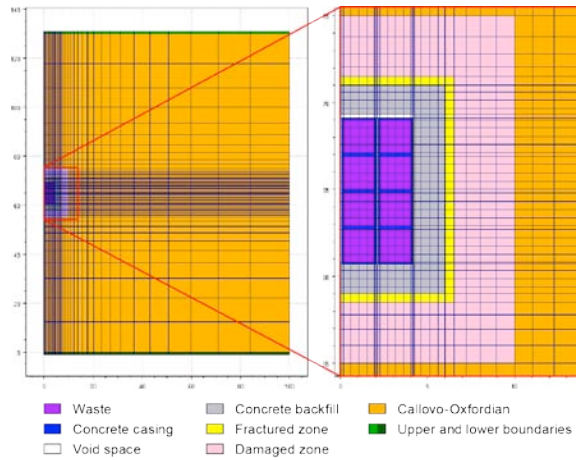


Figure 3. The 'elementary' Couplex-Gaz 1b model.

Preliminary tests (Hubschwerlen et al., 2012) showed a threshold for the parallel acceleration, in this case for as little as around 800 unknowns per parallel process. However, acceleration only became significant and approached linear rates above 2500 unknowns per parallel process. Therefore, and in order to make tests with the widest possible spectrum of parallel processes, the test model was extended to around 300 000 unknowns, which was a priori estimated as sufficient. Indeed, Azurite has 16 nodes with the possibility to run up to 8 processes per node, i.e. a maximum of 128 parallel processes. This was achieved by connecting 64 layers of the reference mesh layout, resulting in a 158,720-element mesh (Figure 4). Some other meshes with intermediate sizes were created for the scalability study upon the same method.

Even with the 2480-element reference mesh, the computation time over the entire simulation period of 50 000 years requires considerable CPU time. After observing the physical behavior of the model, we decided to work with a simulation period of 10 years starting at $t=1000$ years, corresponding to a critical period in which water pressure rises as hydrogen is produced, inducing nonlinear behavior. This reduction enabled CPU times reduced to a few minutes for the test simulations.

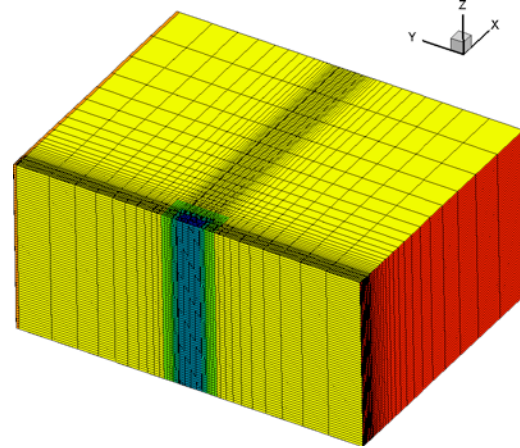


Figure 4. Couplex-Gaz 1b mesh replicated 64 times: 158 720 elements. Stacking is performed in the z dimension.

Couplex-Gaz 1b: Test of the partitioning algorithm

The test of the METIS algorithms was performed as described above. All 16 nodes of the Azurite cluster were used, and the number of tested partitions np was varied from 16 to 128. Results showed no clear trend in terms of computation efficiency (Figure 5). For example, for the number of partitions $np = 96$, the results were good, while for $np = 80$ and $np = 112$, they were poor (for all three algorithms). Reasons for this apparent randomness can be found in the total number of solver iterations performed for the different runs (Hubschwerlen et al., 2012) owing to the conditioning of systems resulting from the domain partitioning.

However, from observing cases with good solver convergence, we could see that the best efficiency for a relatively small number of partitions ($np = 48$) was reached by the METIS *Recursive* partitioning, whereas for larger partitions ($np = 96, 112$ and 128), the best efficiency was reached with *Kway*. This

confirmed the recommendations on the METIS algorithm selection found in Zhang et al. (2008). Moreover, *VKway* performed poorly overall. This algorithm minimizes communication between nodes best, but this does not seem to be a bottleneck for such relatively small partitions on the Azurite cluster, which is equipped with a good Infiniband connection.

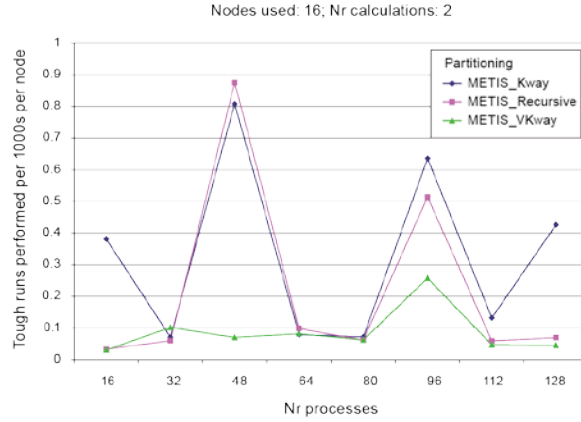


Figure 5. Couplex-Gaz 1b: Efficiency versus number of parallel processes for the test of the METIS partitioning algorithms on Azurite; simulation period 1000-1010 years; 158 720 elements model.

Couplex-Gaz 1b: Search of the optimal parallel process distribution

The search for best distribution of parallel processes on nodes was done via the two described tests: varying the node load on the cluster, and testing the influence of repartitioning the processes of one calculation on the nodes.

Variation of the node load was done according to the methodology. The computation used is a METIS *Recursive* with 8 parallel processes for the 64-layer (158 720 elements) Couplex-Gaz 1b problem. The node load was increased progressively from 1 to 8 processes per node by increasing the number of simultaneously running computations on an 8-node restriction of the Azurite cluster. Under these conditions, the individual run time of a single computation progressively rises with the node load, because each single computation has to share node resources (Table 1). Degradation of performance becomes more important when more than four processes are running simultaneously per node—attributed to saturation of the node’s

cache memory. This hypothesis was confirmed by a supplementary test under the same conditions with a smaller 8-layer (19,840 elements) Couplex-Gaz 1b problem. This lighter problem showed much less degradation of CPU time for the same node loads, because nodes had to process many fewer unknowns simultaneously.

Table 1. Couplex-Gaz 1b – Degradation of pure performance, measured as % increase in CPU time, when increasing the number of processes per node; comparison for 64-layer and 8-layer models runs on 8 nodes of Azurite.

Node load variation	CPU time increase	
	64-layers Couplex-Gaz 1b case	8-layers Couplex-Gaz 1b case
1 to 2 p. p. node	+20%	+0%
2 to 4 p. p. node	+22%	+0%
4 to 8 p. p. node	+90%	+21%
1 to 4 p. p. node	+46%	+0%
1 to 8 p. p. node	+180%	+22%

For the 64-layer mesh, (Figure 6), with up to four parallel processes running per node, the efficiency increased strongly and almost linearly because of the better capacity utilization of the cores in the nodes. Beyond this threshold, efficiency stagnated, because the nodes’s cache memory saturation acted as a bottleneck.

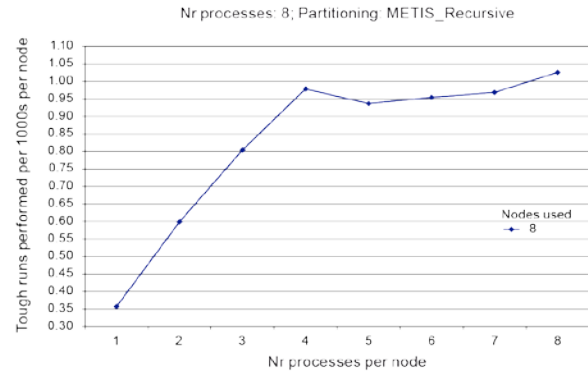


Figure 6. Couplex-Gaz 1b: Efficiency as a function of the number of processes simultaneously assigned to each of the 8 Azurite nodes; 158 720 elements model.

Repartition of the processes for a single calculation on the Azurite cluster nodes had no significant influence, as shown by results in

Figure 7. This means the communication between nodes is not a performance bottleneck for partition domains of this size (average 40 000 unknowns) on this particular cluster.

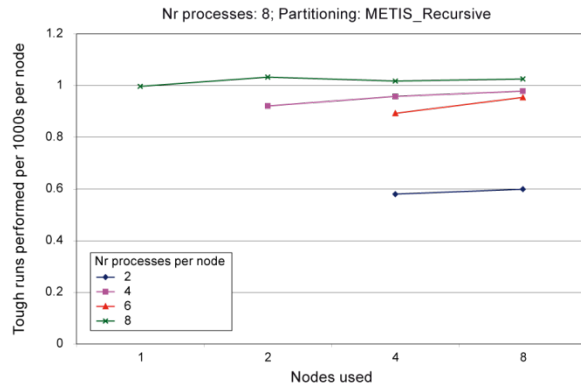


Figure 7. Couplex-Gaz 1b: Efficiency as a function of the number of Azurite nodes involved for the calculation, for different constant node loads; 158 720 elements model.

Couplex-Gaz 1b: Study of scalability

Scalability testing was carried out with the METIS-*Kway* partitioning algorithm, since the behavior of TOUGH2-MP for large numbers of partitions is of most interest. The number of processes was varied from 2 to 128 (i.e., the maximal possible on Azurite), and node load was ensured constant to 8 processes running per node, if necessary by running several identical runs simultaneously (for low numbers of parallel processes). The test was repeated for six different multiplication factors of the reference Couplex-Gaz 1b mesh with 4960 (2 layers) to 158 720 elements (64 layers).

Analysis of the measured efficiencies when increasing the number of parallel processes (Figure 8) shows two phases:

- Increase in the efficiency until a peak, corresponding to supralinear acceleration. This phase is not encountered for the smallest mesh, since partition size is too small.
- Decrease in efficiency. At first, this decrease corresponds to sublinear parallel acceleration when CPU gains on one run do not compensate for the extra usage of cluster resources. Over a certain number of domains, the number of unknowns per parallel process becomes too small to allow

any parallel acceleration. Individual CPU times even increase because the supplementary communication overhead between subdomains becomes a bottleneck.

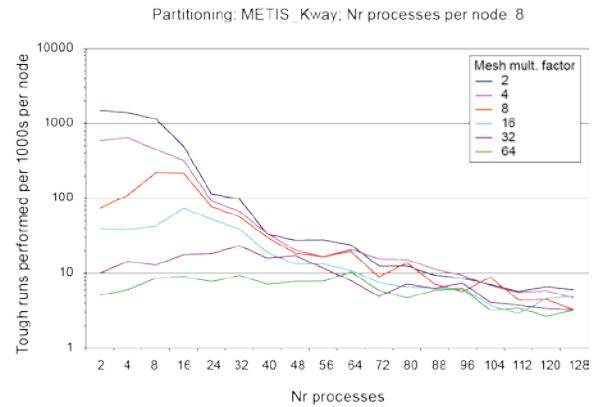


Figure 8. TOUGH2-MP scalability for the Couplex-Gaz 1b: Efficiency on a logarithmic scale as a function of number of parallel processes for various mesh sizes.

The bigger the meshes, the later the optimal efficiency peak and decrease occur. The measured number of domain partitions achieving best efficiency is proportional to the mesh size (Figure 9), resulting in a constant optimal partition size of about 5000 unknowns and showing good scalability of TOUGH2-MP.

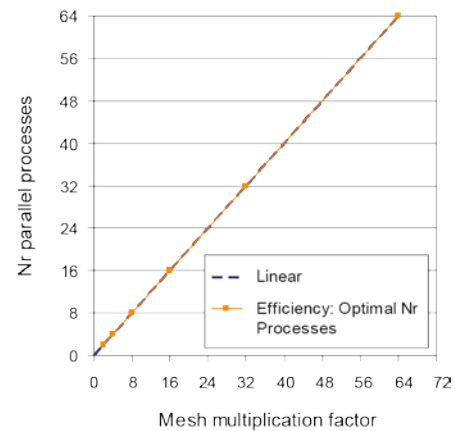


Figure 9. TOUGH2-MP scalability: for each multiplication factor of the Couplex-Gaz 1b mesh (which is proportional to number of unknowns), number of parallel processes reaching peak efficiency.

DIRECT APPLICATION TO A LARGE CASE

The MAVL 3D case

The selected large case is the MAVL 3D model used to calculate the coupled hydrogen production and heat generation arising in a radioactive waste disposal drift containing ILW-LL (Poller et al., 2009). This 55 345 elements 3D mesh (Figure 10) was run with TOUGH2-MP EOS5 with nonisothermal conditions (three equations per element). It leads to a 166 035 equations problem.

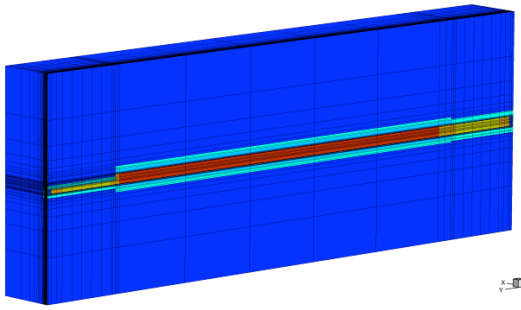


Figure 10. Mesh used for the MAVL 3D case.

To obtain a test case that could be run in a reasonable time, the simulation was reduced to a 0–50 years period, which according to Figure 11, corresponds to a phase in which the strongest variations in pressure, saturation degree, and temperature occur. Accordingly, the CPU time is reduced from several hours to approximately 600 seconds with 8 parallel processes.

Testing of the different partitioning algorithms showed that best efficiency is achieved with *Kway* and 16 processes (Figure 12). The size of one partition is then close to 10,000 unknowns.

Node load performed with 8 processes on 8 nodes showed results comparable to those of the Couplex-Gaz 1b case. Computational efficiency steeply increases until 4 processes per node. Over this threshold, efficiency improvement continues, but is less pronounced. Different process repartitions over several nodes of the cluster have little influence as well (Figure 13).

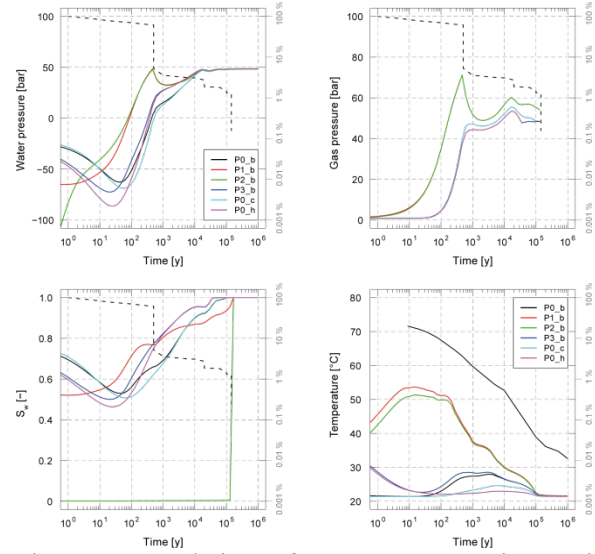


Figure 11. Evolution of pressure, saturation and temperature over time for the reference MAVL 3D computation.

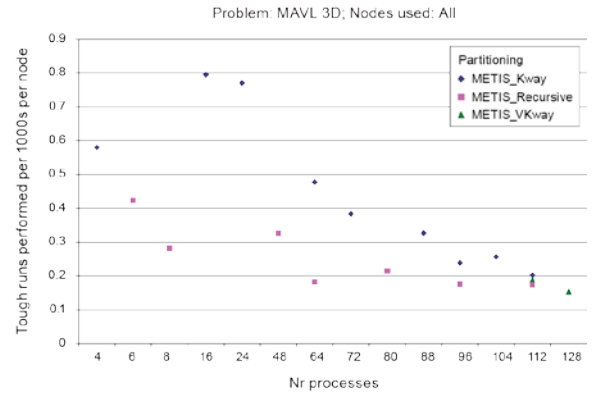


Figure 12. MAVL 3D – Efficiency as a function of the number of parallel processes for the 3 available METIS partitioning algorithms.

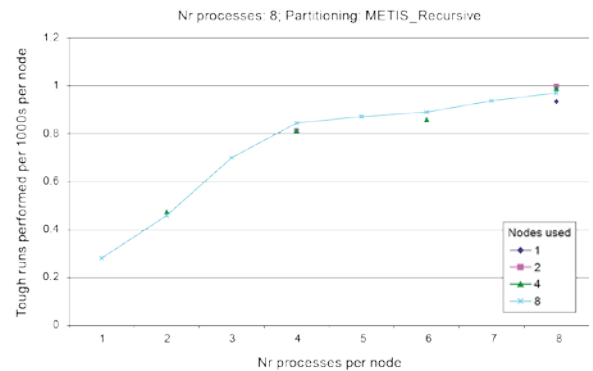


Figure 13. MAVL 3D – Efficiency expressed in terms of performed runs per time per Azurite node as a function of the node load.

Comparison between large scale results

The results obtained with the largest Couplex-Gaz 1b problem (64-layer mesh) and the MAVL 3D case lead to comparable findings regarding the parallel settings used to reach optimal performance. In the Couplex-Gaz 1b case, best efficiency was achieved with $np = 64$, and results with $np = 16$ and $np = 32$ are very good as well. This corresponds to partition sizes with average number of unknowns in the 5000–20,000 range. Best efficiency with the MAVL case is obtained for $np = 16$ (10,000 unknowns per partition) and $np = 24$ (7000 unknowns per partition). Thus, both models are processed optimally with similarly defined partitions.

CONCLUSIONS

In this paper, we presented a methodology to improve the conditions for use of TOUGH2-MP on a cluster. Although predicting with precision the best parallel settings to run a given simulation is impossible—because of the heterogeneous number of solver iterations performed by TOUGH2-MP with different partitions—it is possible to derive some good practices from the experience gathered on Azurite. First, the METIS *Recursive* algorithm is preferable for small numbers of partitions, and *Kway* is preferable for larger numbers of partitions. Second, efficiency is optimal when cluster nodes are well loaded, which corresponds for Azurite to at least 4 parallel processes per node with 10,000 unknown partitions. Maximal efficiency seems to be reached for partition domains of about this size, and is not influenced by the way the different parallel processes are distributed on the cluster nodes. Finally, both tests with Couplex-Gaz 1b and MAVL 3D models have demonstrated good scalability with TOUGH2-MP. This scalability allows using the methodology on a small-scale representative model in order to prepare a sensible parallel setup for a large-scale simulation.

REFERENCES

- Andra, *Synthesis. Evaluation of the feasibility of a geological repository in an argillaceous formation. Dossier Argile 2005. Collection les Rapports*, Andra, Châtenay-Malabry, France, 2005.
- Andra, *Cas test Couplex-Gaz 1 : modélisation 2D d'une alvéole de déchets de moyenne activité à vie longue*. http://www.andra.fr/couplex/Exercice_Couplex_Gaz_1.pdf, 2006. [Accessed 2 July 2012]
- Hubschwerlen, N., K. Zhang, G. Mayer, J. Roger, and B. Vialay, Using TOUGH2-MP on a cluster - optimization methodology and study of scalability, *Computers & Geosciences*, doi:10.1016/j.cageo.2012.03.005, 2012.
- Karypis, G. and V. Kumar, *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices, V4.0*, Technical Report, Department of Computer Science, University of Minnesota. 1998.
- Karypis, G. and V. Kumar, A Fast and Highly Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing*, Vol. 20, No. 1, 359-392, 1999
- Poller, A., C. P. Enssle, G. Mayer, and J. Wendling, Repository-scale modeling of the long-term hydraulic perturbation induced by gas and heat generation in a geological repository for high and intermediate level radioactive waste— Methodology and results. *TOUGH Symposium 2009, Lawrence Berkeley National Laboratory, Berkeley, California*, 2009.
- Pruess, K., C. Oldenburg, and G. Moridis, *TOUGH2 User's Guide, Version 2.0*, Report LBNL-43134, Lawrence Berkeley National Laboratory, Berkeley, Calif., U.S.A., 1999.
- Zhang, K., Y. S. Wu, and K. Pruess, *User's guide for TOUGH2-MP – A Massively parallel Version of the TOUGH2 Code –*, Report LBNL-315E, Lawrence Berkeley National Laboratory, Berkeley, Calif., 2008.